Selecting relevant predictors: impact of variable selection on model performance, uncertainty and applicability of models in environmental decision making

Gert Everaert, Javier E. Holguin, Peter L.M. Goethals
Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, 9000 Ghent, Belgium.
(Gert.Everaert@UGent.be)

Abstract: One of the crucial steps when developing models is the selection of appropriate variables. In this research we assessed the impact variable selection on the model performance and model applicability. Regression trees were built to understand the relationship between the ecological water quality and the physical-chemical and hydromorphological variables. Different model parameterizations and three combinations of explanatory variables were used for developing the trees. Once constructed, they were integrated with the water quality model (PEGASE) and used to simulate the future ecological water quality. These simulations were summarized per combination of explanatory variables and compared.
Three key messages summarize our conclusions. First, it was confirmed that different parameterizations alter the statistical reliability of the trees produced. Secondly, it was found that statistical reliability of the models remained stable when different combinations of explanatory variables were implemented. The determination coefficient (R^2) ranged from 0.68 to 0.86; Kappa statistic (K) ranged from 0.15 and 0.46; and the percentage of Correctly Classified Instances (CCI) from 33 to 59%. Thirdly, when applying the models on an independent dataset consisting of future physical-chemical water quality data, different conclusions may be taken, depending on the combination of variables used.

***Keywords**:* model applicability; model performance; regression trees; variable selection

1 INTRODUCTION

Ecological models have been often used in environmental decision making (e.g. Argent et al., 2009; Mouton et al., 2009). Several guidelines on the model development have been written (e.g. Zuur et al., 2010). These guidelines assist researchers when taking decisions during the model development process. For instance, data preprocessing and model parameterization are two key aspects to obtain a reliable and applicable model (Everaert & Goethals, submitted). Similarly, an important consideration is which and how many explanatory variables should be included to make valid predictions because the selection of relevant variables affects the models produced and their statistical reliability (Elith & Leathwick,

2009). Consequently, it is expected that the applicability and the conclusions drawn, may change depending on the explanatory variables used. If too few variables are included not all variance can be explained, whereas including too many variables results in complex, over-trained models, not suitable to be applied in environmental decision making.
In this research the impact of different variable combinations on the statistical reliability and applicability was illustrated. First, the dataset was explored and pre-processed. Next, three possible combinations of explanatory variables were defined after which regression trees were developed. In a last step, the models inferred were applied on the future water quality predictions for three target years.

2 MATERIALS AND METHODS

2.1 Background and description of the dataset

The dataset was compiled in the scope of a project performed at the Laboratory of Environmental Toxicology and Aquatic Ecology (Ghent University, Belgium) on the authority of the Flemish Environment Agency (VMM; Figure 1). The main objective of the project was to develop predictive models relating physical-chemical and ecological surface water conditions and, in the end, to help river manager decide where to allocate their limited resources for river restoration. Data needed to develop such models encompass physical-chemical, hydromorphological and biological quality data. More information on the project and the corresponding conclusions are described by Everaert et al. (2010) and Pauwels et al. (2010).

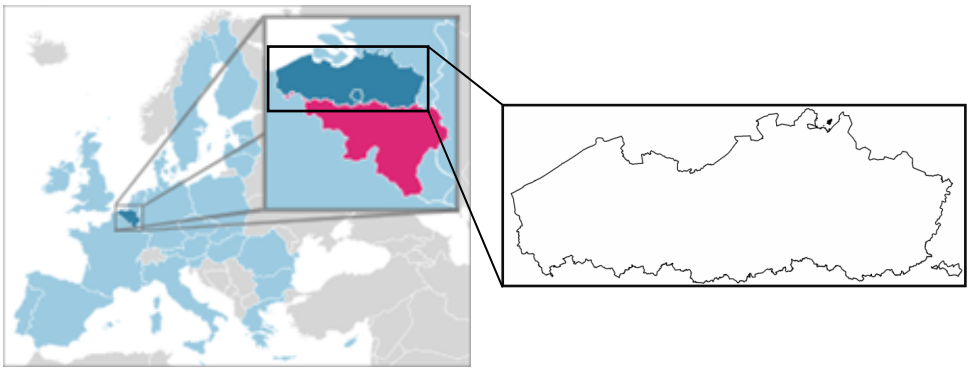

Figure 1. Location of Flanders in Belgium, Europe

The physical-chemical variables were available in the form of statistical derivatives over one year (mean, median, minimum, maximum and 5% - 10% - 90% - 95% percentiles). The same statistical derivatives per physical-chemical variable were used as proposed by Schneiders et al. (2009): maximum Biological Oxygen Demand (BOD_5, mg O_2/L), maximum Chemical Oxygen Demand (COD, mg O_2/L), median Kjeldahl nitrogen concentration (KjN, mg N/L), median nitrate concentration (NO_3^--N, mg N/L), minimum oxygen concentration (DO, mg O_2/L), average orthophosphate concentration (PO_4^{3-}-P, mg P/L) and average total phosphorous concentration (Pt, mg P/L). All substances were analyzed in accordance to the standards of ISO 17025.
The mean slope of the watercourse was used to quantify one aspect of the hydromorphology of the sampling locations. The method assumed that the altitude

of a watercourse, averaged over a certain distance, is a reasonable estimator of the slope of a watercourse and is related to the flow velocity (Dumortier et al., 2009).
The MMIF (Multimetric Macroinvertebrate Index Flanders), ranging from 0 to 1, was used to quantify the ecological water quality of the Flemish water courses. The method to assess the ecological status of Flemish surface waters based on the macroinvertebrate community is discussed in detail by Gabriels et al. (2010). In the context of the European Water Framework Directive (WFD) and for transparency towards decision makers, the MMIF is converted to five ecological quality classes ("bad, "poor", "moderate", "good" and "high"). The quality classes "good" and "high" were aggregated in one class, named "good_high" as limited records were available for the best water quality class.

2.2 Data pre-processing

Physical-chemical, hydromorphological and biological quality data were combined based on location and year of sampling. This resulted in an unprocessed dataset of 1716 samples. Subsequently, only complete cases were retained, outliers were removed and the dataset was stratified for the response variable by means of subsampling (Araujo and Guisan, 2006; Everaert & Goethals, submitted). For the stratification, in each quality class as many samples were randomly selected as available in the least represented quality class. A summary of the stratified dataset, containing 240 out of 1716 cases, can be found in Table 1. The Pearson correlation coefficients were calculated to explore the relations between the variables available (Table 2).

Table 1. Observed characteristics in the Flemish watercourses, based on 240 records.

Variable	Statistical derivative	Unit	Minimum	Maximum	Mean
MMIF	-		0	1	0.5
Slope	mean	m/km	0	10.3	0.9
BOD_5	maximum	mg/l	0	60.3	7.8
COD	maximum	mg/l	13	299	58
KjN	median	mg N/l	0	11.0	2.4
NO_3^--N	median	mg N/l	0	9.5	3.6
DO	minimum	mg/l	0.4	7.7	4.2
PO_4^{3-}-P	mean	mg P/l	0	2.1	0.3
Pt	mean	mg P/l	0	2.8	0.7
MMIF class	Bad	Poor	Moderate	Good_high	
# samples	60	60	60	60	

2.3 Model building, validation and simulation

Regression trees were built through applying the R package rpart (R Development Core Team, 2009). Rules relating the MMIF with physical-chemical and hydromorphological conditions were created using the Classification and Regression Trees (CART) algorithm (Breiman et al., 1984). Regression models

were produced for multiple settings considering the pruning level and the minimum number of records per leaf (min.objects). The pruning level varied from 0.01 to 0.18 and the minimum number of records from 2 to 10. Pairplots (Zuur et al., 2009) give insight in the impact of the different model development settings on the predictive power.

Three-fold cross-validation was implemented to train and validate the regression model. The models were evaluated based on the determination coefficient (R^2), the percentage of Correctly Classified Instances (CCI) and Kappa statistic (K). The CCI was calculated as the percentage of true positive and true negative predictions. The K measured the percentage of true positive and true negative predictions, but adjusted these values for the amount of agreement that could be expected due to randomness (Cohen, 1960; Fielding and Bell, 1997). Values for R^2 and K range from 0 to 1 and a value close to 1 indicates a better model prediction. In order to have a satisfactory model performance, the CCI and K value should reach at least 70% and 0.4 respectively (Gabriels et al., 2007).

Table 2. Pearson correlation coefficients based on 240 records. Correlated variables are highlighted in bold.

	MMIF	**Slope**	**BOD_5**	**COD**	**KjN**	**NO_3^--N**	**DO**	**PO_4^{3-}-P**	**Pt**
MMIF	1.00	0.14	-0.48	-0.31	**-0.65**	-0.24	0.57	**-0.63**	**-0.64**
Slope		1.00	0.03	-0.02	-0.09	0.08	0.18	0.01	0.01
BOD_5			1.00	0.59	0.46	0.05	-0.49	0.50	0.53
COD				1.00	0.27	-0.03	-0.34	0.36	0.42
KjN					1.00	0.08	**-0.60**	**0.74**	**0.73**
NO_3^--N						1.00	0.03	0.13	0.04
DO							1.00	-0.55	-0.59
PO_4^{3-}-P								1.00	**0.88**
Pt									1.00

Once the optimal model parameterization range was found, regression trees were developed for three combinations of explanatory variables (Table 3). A first option was to use all variables available (Table 1). In a second approach only the non-correlated explanatory variables were used (Table 2). Correlated variables were dropped based on the correlation coefficient. An alternative consideration to detect collinearity is a Principle component Analysis (PCA) (Zuur et al., 2010). However, for simplicity we opted for the correlation coefficient. In the final approach variables were selected based on expert knowledge (Table 3).

Table 3. Variables used to develop regression models

	Predictors used
Approach A	Slope, BOD_5, COD, KjN, NO_3^--N, DO, PO_4^{3-}-P, Pt
Approach B	Slope, BOD_5, KjN, NO_3^--N, DO
Approach C	Slope, KjN, DO

Per variable combination three regression models were constructed (pruning level was 0.02, 0.04 and 0.06). Subsequently, each of those models was implemented on future physical-chemical water quality conditions simulated via the water quality model 'Planification Et Gestion de l'ASsainissement des Eaux' (PEGASE; Deliege et al., 2010). The PEGASE-model simulates the physical-chemical water conditions for three target years: 2006, 2015 and 2027. So, the three regression trees that

were produced per variable combination were applied on the PEGASE-simulations. This resulted in predictions for the future ecological water quality. Finally, per variable combination and per PEGASE-target year, the average future ecological water qualities were calculated and visualized.

3 RESULTS AND DISCUSSION

3.1 Model parameterization and variable selection related to model performance

R^2, CCI and Kappa statistic were, not surprisingly, positively correlated (Figure 2A). At increasing pruning levels, simpler models were generated, but also reliabilities shrunk. Low pruning levels often resulted in complex trees, with better modelling performances (Figure 2A). The influence of the minimum number of observations per leaf was limited; performance criteria did not change with varying values (Figure 2A). Similar conclusions were drawn by Everaert & Goethals (submitted).
Interestingly, the variable selection did not influence the model performances. The predictive performance remained stable over the variable combinations. The only variation noticed, was related to the different model parameterization. Regardless the variable combination, the R^2 ranged from 0.68 to 0.86; K ranged from 0.15 and 0.46; and the CCI from 0.33 to 0.59 (Figures 2A, 2B and 2C).

3.2 Model parameterization and variable selection related to model applicability

In the previous paragraphs, it was described that the variable selection did not influence the model performance. However, when applying these models to future water quality simulations, variable selection did influence the model applicability. Different predictions were found per variable combination (Figure 3).
From the three variable combinations it was possible to derive that the ecological water quality will improve from 2006 to 2027. However, the more variables involved, the more distinct the improvement was, and the better the improvement could be visualized (Figure 3).

It was already mentioned multiple times that the selection of adequate predictors is essential to make reliable models (Elith & Leathwick, 2009). Interestingly, model performances remained stable with changing variable combinations (Figures 1A, 1B and 1C). However, when applying the models, it was obvious that different water quality predictions were found per variable combination. This conclusion is related to recent work by Everaert & Goethals (submitted); other quality aspects than statistical reliability are equally important in the model selection.
The relative low statistical reliabilities of the models produced may indicate that the number of variables that were included in the regression model were insufficient. However, to date the PEGASE-model only predicts a limited number of variables. Therefore, it is recommended to include additional integrative variables (e.g. conductivity and hydromorphological variables) in the water quality simulations in order to optimize the predictive power of the models.

4 CONCLUSIONS

In this research we confirmed that model performances and model applicability are altered by the model parameterization. We found that model performances remained stable when different combinations of explanatory variables were used. However, when applying those models on an independent dataset, different conclusions and decisions may be taken, depending on the combination of variables used.

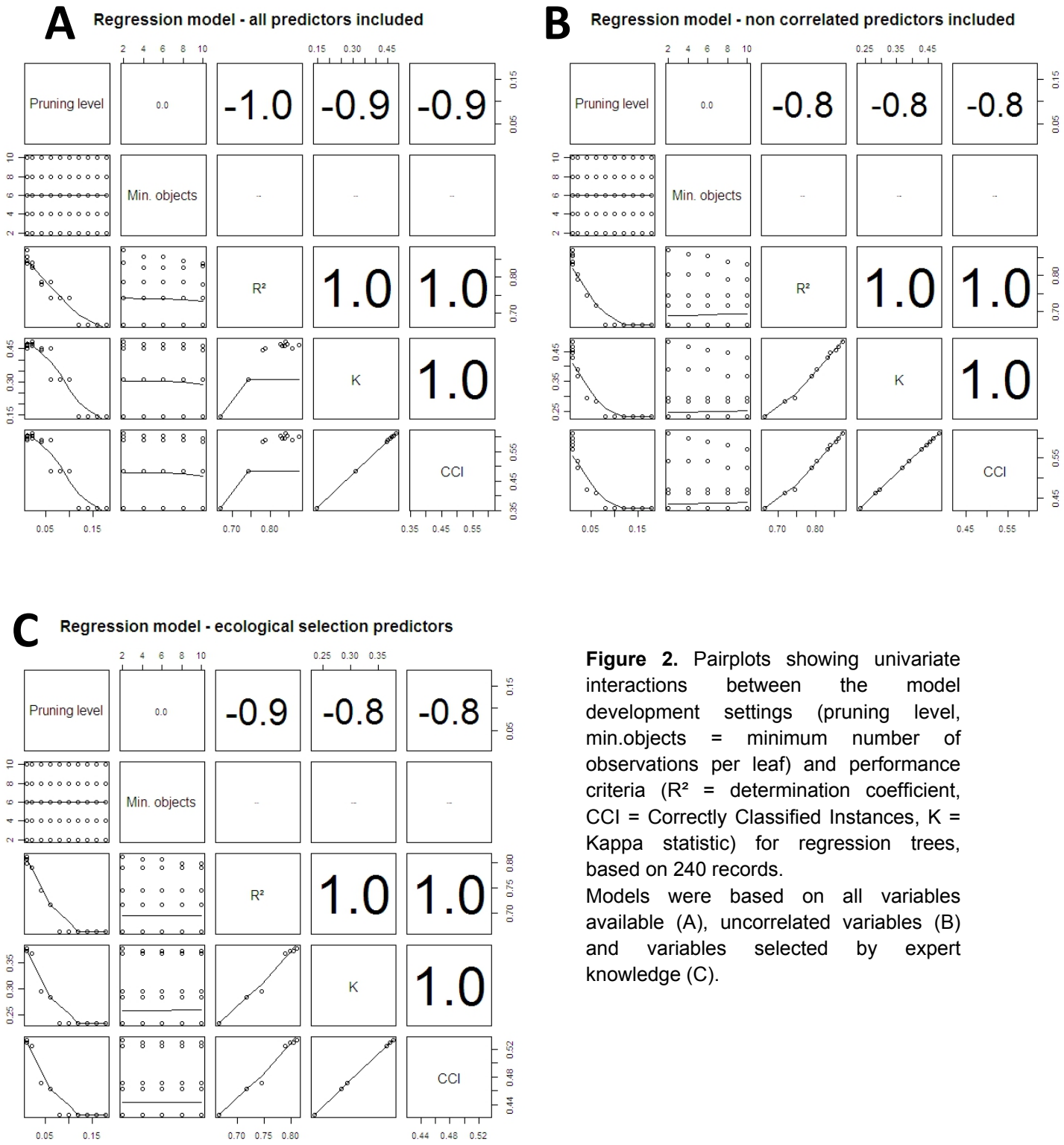

Figure 2. Pairplots showing univariate interactions between the model development settings (pruning level, min.objects = minimum number of observations per leaf) and performance criteria (R^2 = determination coefficient, CCI = Correctly Classified Instances, K = Kappa statistic) for regression trees, based on 240 records.
Models were based on all variables available (A), uncorrelated variables (B) and variables selected by expert knowledge (C).

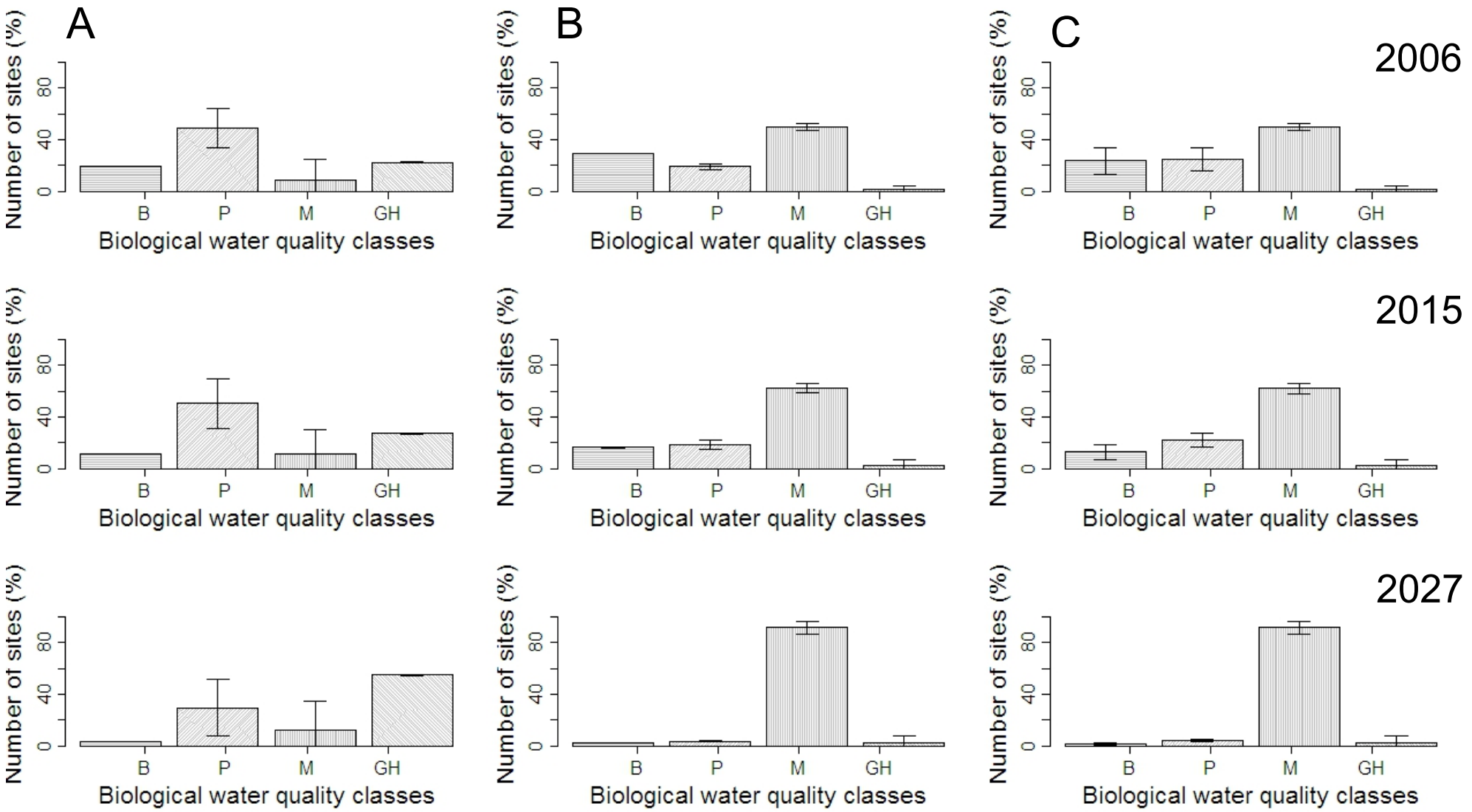

Figure 3. Visualization of the ecological water quality predictions generated by applying the regression tree on the PEGASE-simulations for 2006, 2015 and 2027. Models were based on all variables available (A), uncorrelated variables (B) and, variables selected by expert knowledge (C). Water quality predictions are subdivided in four ecological quality classes: bad (B, horizontal bars), poor (P, 45° bars) , moderate (M, vertical bars), and good_high (GH, 135° bars) .

ACKNOWLEDGMENTS

Javier E. Holguin is currently supported by a doctoral fellowship from the Special Research Fund of Ghent University (BOF) in Belgium. We would like to thank the Flemish Environment Agency (VMM) for the use of their data.

REFERENCES

Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography,* 33, 1677-1688.

Argent, R.M., Perraud, J.M., Rahman, J.M., Grayson, R.B., Podger, G.M., 2009. A new approach to water quality modelling and environmental decision support systems. *Environmental Modelling & Software,* 24, 809-818.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, USA, 358 pp.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.

Deliege, J.F., Everbecq, E., Magermans, P., Grard, A., Bourouag, T., Blockx, C., Smitz, J., 2010. PEGASE, an integrated river/basin model dedicated to surface water quality assessment: application to cocaine. *Acta Clinica Belgica,* 65, 42-48.

Dumortier, M., De Bruyn, L., Hens, M., Peymen, J., Schneiders, A., Van Daele, T., Van Reeth, W., 2009. Natuurverkenning 2030. Natuurrapport Vlaanderen, NARA 2009. Mededeling van het Instituut voor Natuur- en Bosonderzoek, INBO,Brussel, 224 pp. (in Dutch).

Elith, J. and Leathwick, J.R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. - In: Annual Review of Ecology Evolution and Systematics. *Annual Reviews*, pp. 677-697.

Everaert, G., Goethals, P.L.M. Searching for a representative cluster of trees in an unexplored forest: impact of data stratification, cross-validation and pruning on classification and regression trees' reliability. Submitted to Environmental Modelling & Software.

Everaert, G., Pauwels, I.S., Goethals, P.L.M., 2010. Development of data-driven models for the assessment of macroinvertebrates in rivers in Flanders, In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), 5th Biennial meeting of the International Congress on Environmental Modelling and Software (iEMSs 2010): Modelling for environment's sake International Environmental Modelling and Software Society (iEMSs) Ottawa, ON, Canada.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation,* 24, 38-49.

Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. *Aquatic Ecology*, 41, 427-441.

Gabriels, W., Lock, K., De Pauw, N., Goethals, P.L.M., 2010. Multimetric Macroinvertebrate Index Flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium). *Limnologica*, 40, 199-207.

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2009. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. *Environmental Modelling & Software,* 24, 982-993.

Pauwels, I.S., Everaert, G., Goethals, P.L.M., 2010. Integrated river assessment by coupling water quality and ecological assessment models In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), 5th Biennial meeting of the International Congress on Environmental Modelling and Software (iEMSs 2010) : Modelling for environment's sake International Environmental Modelling and Software Society (iEMSs) Ottawa, ON, Canada.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Schneiders, A., Simoens, I., Belpaire, C., 2009. Waterkwaliteitscriteria opstellen voor vissen in Vlaanderen. Rapporten van het Instituut voor Natuur- en Bosonderzoek. INBO, Brussel, 94 pp. (in Dutch).

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1, 3-14.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer Science+Business Media, LLC 2009, New York.